EE E6820: Speech & Audio Processing & Recognition

Lecture 4: Auditory Perception

Mike Mandel <mim@ee.columbia.edu>
Dan Ellis <dpwe@ee.columbia.edu>

Columbia University Dept. of Electrical Engineering
http://www.ee.columbia.edu/∼dpwe/e6820

February 10, 2009

1. Motivation: Why & how
2. Auditory physiology
3. Psychophysics: Detection & discrimination
4. Pitch perception
5. Speech perception
6. Auditory organization & Scene analysis

Outline

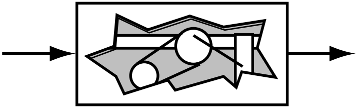
Why study perception?

- Perception is messy: can we avoid it?
  No!
- Audition provides the 'ground truth' in audio
  - what is relevant and irrelevant
  - subjective importance of distortion (coding etc.)
  - (there could be other information in sound...)
- Some sounds are 'designed' for audition
  - co-evolution of speech and hearing
- The auditory system is very successful
  - we would do extremely well to duplicate it
- We are now able to model complex systems
  - faster computers, bigger memories

How to study perception?

Three different approaches:

- Analyze the example: physiology



  - dissection & nerve recordings

- Black box input/output: psychophysics

  - fit simple models of simple functions

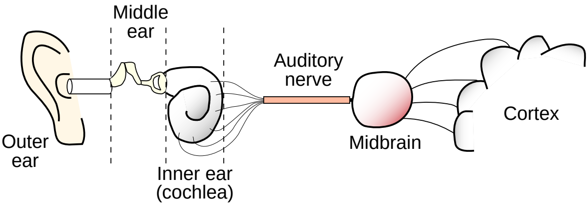
- Information processing models
  - investigate and model complex functions

  e.g. scene analysis, speech perception

Outline

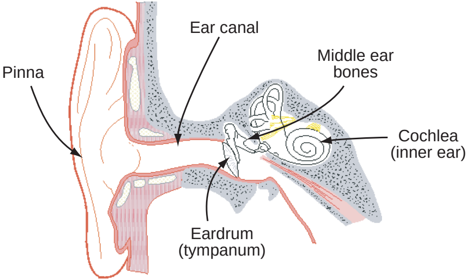
Physiology

- Processing chain from air to brain:



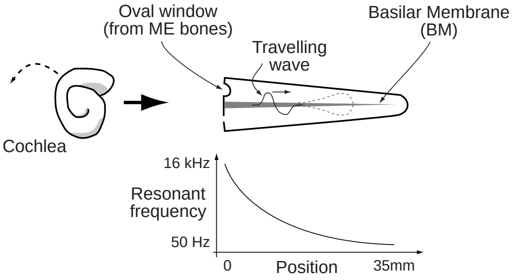
- Study via:
  - anatomy
  - nerve recordings
- Signals flow in both directions

Outer & middle ear



- Pinna ‘horn’
  - complex reflections give spatial (elevation) cues
- Ear canal
  - acoustic tube
- Middle ear
  - bones provide impedance matching

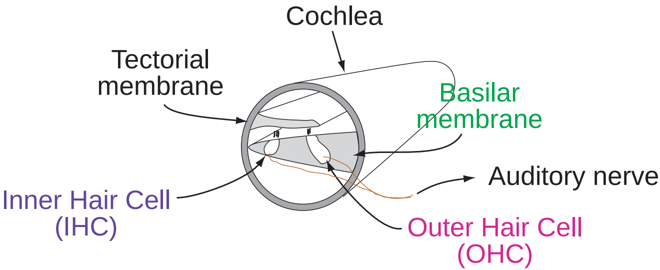
Inner ear: Cochlea



- Mechanical input from middle ear starts traveling wave moving down Basilar membrane
- Varying stiffness and mass of BM results in continuous variation of resonant frequency
- At resonance, traveling wave energy is dissipated in BM vibration
  - Frequency (Fourier) analysis

Cochlea hair cells

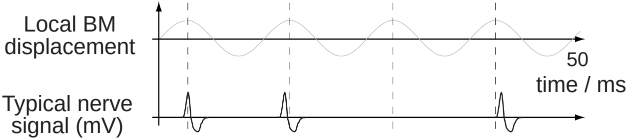
- Ear converts sound to BM motion
  - each point on BM corresponds to a frequency



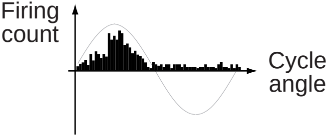
- Hair cells on BM convert motion into nerve impulses (firings)
- Inner Hair Cells detect motion
- Outer Hair Cells? Variable damping?

Inner Hair Cells

- IHCs convert BM vibration into nerve firings
- Human ear has ~3500 IHCs
    - each IHC has ~7 connections to Auditory Nerve
- Each nerve fires (sometimes) near peak displacement



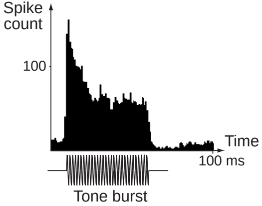
- Histogram to get firing probability



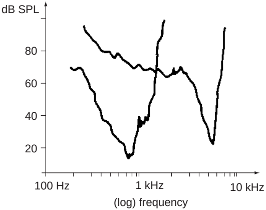
Auditory nerve (AN) signals

Single nerve measurements

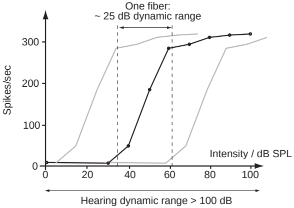
Tone burst histogram



Frequency threshold



Rate vs intensity

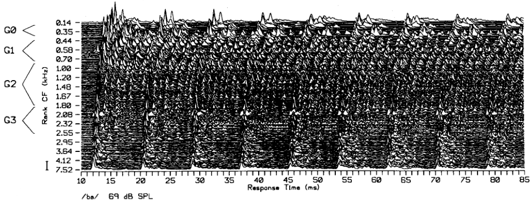


Hard to measure: probe living ANs?

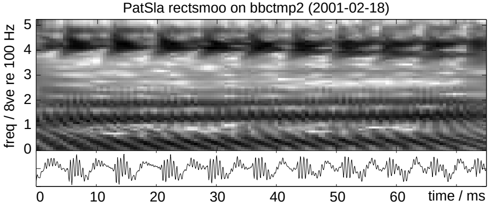
AN population response

All the information the brain has about sound

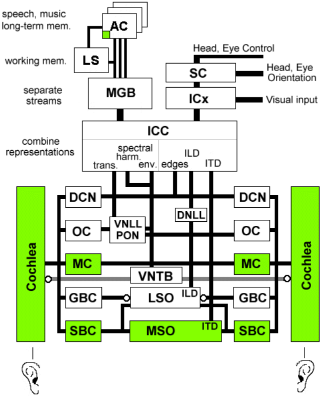
- average rate & spike timings on 30,000 fibers



Not unlike a (constant-Q) spectrogram

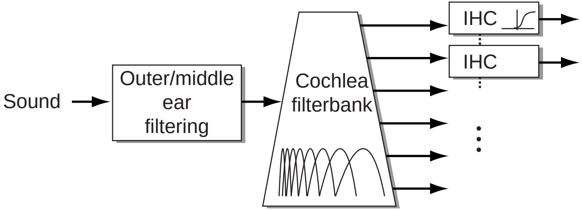


Beyond the auditory nerve

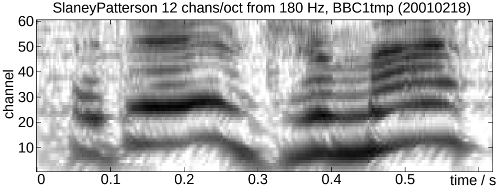


- Ascending and descending
- Tonotopic × ?
  - modulation, position, source??

Periphery models



- Modeled aspects
  - outer / middle ear
  - hair cell transduction
  - cochlea filtering
  - efferent feedback?



Results: ‘neurogram’ / ‘cochleagram’

Outline

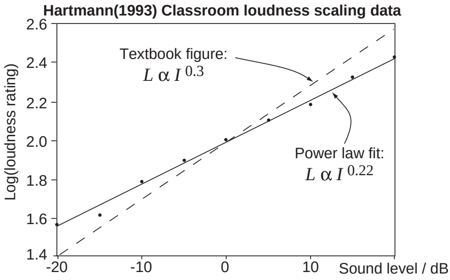
1. Motivation: Why & how
2. Auditory physiology
3. **Psychophysics: Detection & discrimination**
4. Pitch perception
5. Speech perception
6. Auditory organization & Scene analysis

Psychophysics

- Physiology looks at the implementation
  Psychology looks at the function/behavior
- Analyze audition as signal detection: $p(\theta \,|\, x)$
  - psychological tests reflect internal decisions
  - assume optimal decision process
  - infer nature of internal representations, noise, ...
  - $\longrightarrow$ lower bounds on more complex functions
- Different aspects to measure
  - time, frequency, intensity
  - tones, complexes, noise
  - binaural
  - pitch, detuning

Basic psychophysics

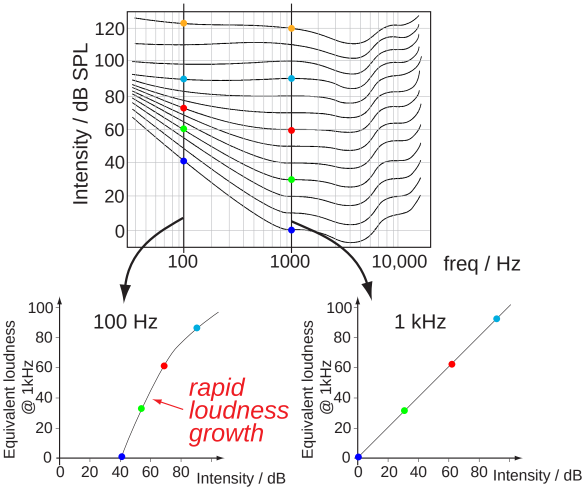
- Relate physical and perceptual variables
  e.g. intensity → loudness
  frequency → pitch
- Methodology: subject tests
  - just noticeable difference (JND)
  - magnitude scaling *e.g.* "adjust to twice as loud"
- Results for Intensity vs Loudness:
  Weber's law $\Delta I \propto I \Rightarrow \log(L) = k \log(I)$



$$\log_2(L) = 0.3 \log_2(I) = 0.3 \frac{\log_{10} I}{\log_{10} 2} = \frac{0.3}{\log_{10} 2} \frac{\text{dB}}{10} = \text{dB}/10$$

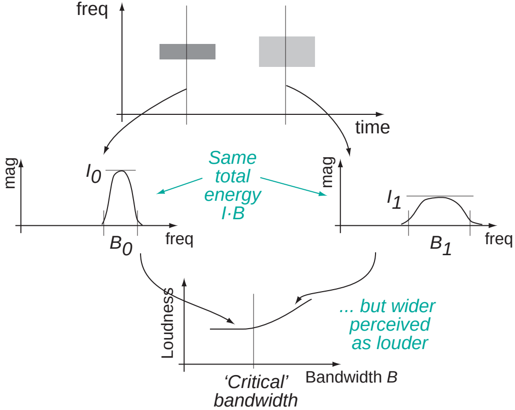
Loudness as a function of frequency

Fletcher-Munsen equal-loudness curves



Loudness as a function of bandwidth

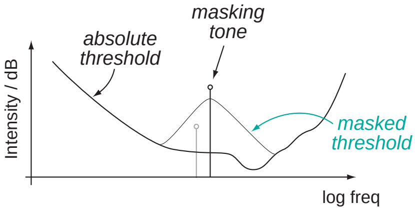
- Same total energy, different distribution
  e.g. 2 channels at −6 dB (not −10 dB)



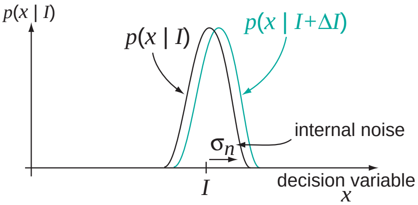
- Critical bands: independent frequency channels
  - ∼25 total (4-6 / octave)

Simultaneous masking

A louder tone can 'mask' the perception of a second tone nearby in frequency:

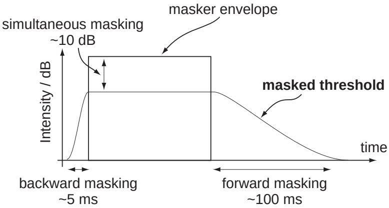


Suggests an 'internal noise' model:

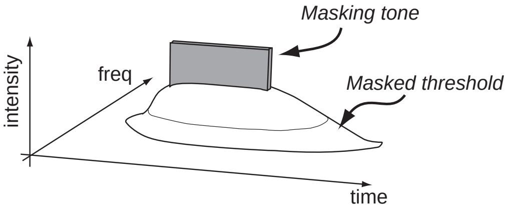


Sequential masking

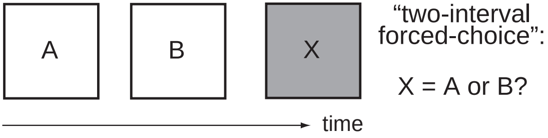
Backward/forward in time:



→ Time-frequency masking 'skirt':



What we do and don't hear

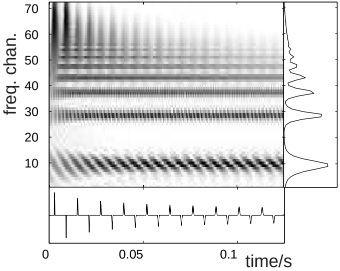


- Timing: 2 ms attack resolution, 20 ms discrimination
  - but: spectral splatter
- Tuning: ~1% discrimination
  - but: beats
- Spectrum: profile changes, formants
  - variables time-frequency resolution
- Harmonic phase?
- Noisy signals & texture
- (Trace vs categorical memory)

Outline

Pitch perception: a classic argument in psychophysics

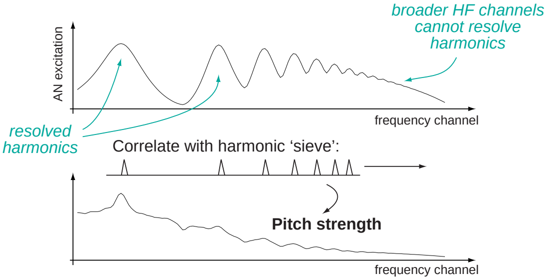
- Harmonic complexes are a pattern on AN



  - but give a *fused* percept (ecological)
- What determines the pitch percept?
  - *not* the fundamental
- How is it computed?
  Two competing models: place and time

Place model of pitch

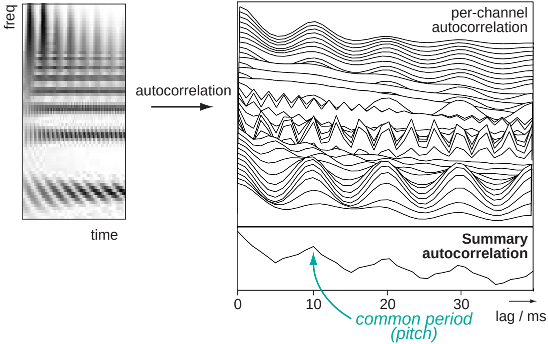
- AN excitation pattern shows individual peaks
- 'Pattern matching' method to find pitch



- Support: Low harmonics are very important
- But: Flat-spectrum noise can carry pitch

Time model of pitch

- Timing information is preserved in AN down to ~1 ms scale
- Extract periodicity by *e.g.* autocorrelation and combine across frequency channels



- But: HF gives weak pitch (in practice)

Alternate & competing cues

- Pitch perception could rely on various cues
  - average excitation pattern
  - summary autocorrelation
  - more complex pattern matching
- Relying on just one cue is brittle
  - *e.g.* missing fundamental

$\longrightarrow$ Perceptual system appears to use a flexible, opportunistic combination

- Optimal detector justification?

$$\begin{aligned}\underset{\theta}{\operatorname{argmax}}\, p(\theta \mid \mathbf{x}) &= \underset{\theta}{\operatorname{argmax}}\, p(\mathbf{x} \mid \theta) p(\theta) \\ &= \underset{\theta}{\operatorname{argmax}}\, p(x_1 \mid \theta) p(x_2 \mid \theta) p(\theta)\end{aligned}$$

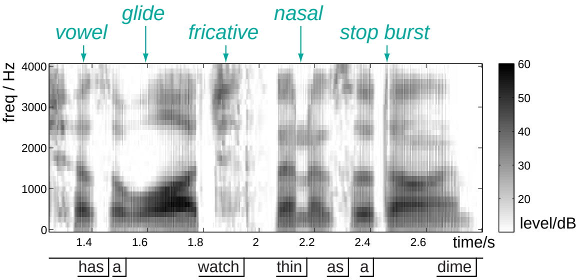
  - if x_1 and x_2 are conditionally independent

Outline

Speech perception

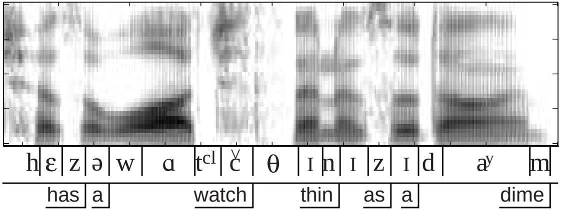
- Highly specialized function
  - subsequent to source organization?

  ... but also can interact
- Kinds of speech sounds

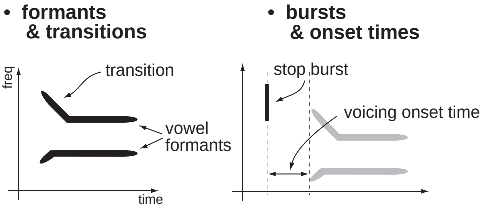


Cues to phoneme perception

Linguists describe speech with phonemes

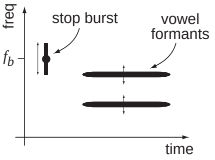


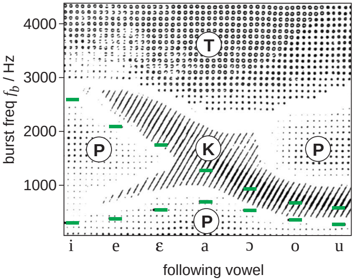
Acoustic-phoneticians describe phonemes by



Categorical perception

- (Some) speech sounds perceived categorically rather than analogically
  - *e.g.* stop-burst and timing:

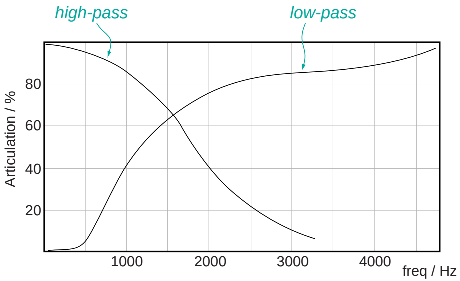




  - tokens within category are hard to distinguish
  - category boundaries are very sharp
- Categories are learned for native tongue
  - "merry" / "Mary" / "marry"

Where is the information in speech?

'Articulation' of high/low-pass filtered speech:



- sums to more than 1...

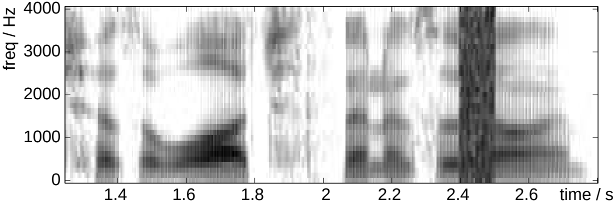
Speech message is highly redundant

e.g. constraints of language, context

→ listeners can understand with very few cues

Top-down influences: Phonemic restoration (Warren, 1970)

What if a noise burst obscures speech?



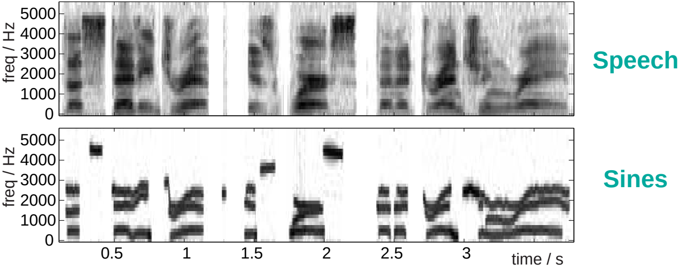
 

- auditory system 'restores' the missing phoneme
  ... based on semantic content
  ... even in retrospect

Subjects are typically unaware of which sounds are restored

A predisposition for speech: Sinewave replicas

Replace each formant with a single sinusoid (Remez et al., 1981)

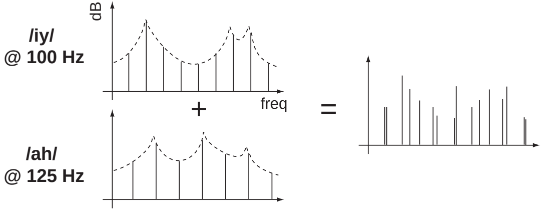


- speech is (somewhat) intelligible
- people hear both whistles and speech ("duplex")
- processed as speech despite un-speech-like

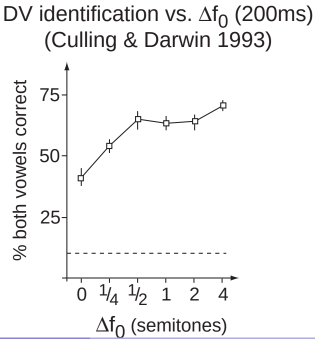
What does it take to be speech?

Simultaneous vowels

Mix synthetic vowels with different f_0s

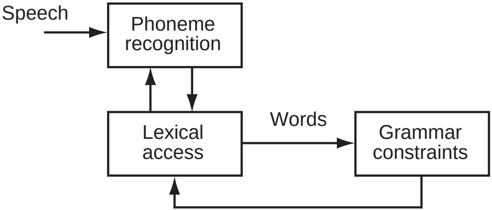


Pitch difference helps
(though not necessarily)



Computational models of speech perception

- Various theoretical-practical models of speech comprehension *e.g.*



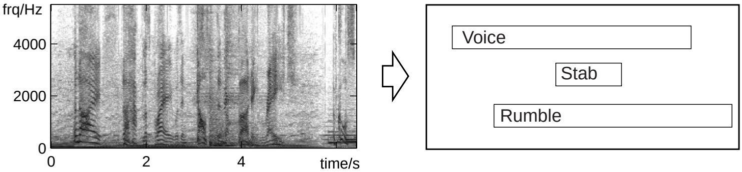
- Open questions:
  - mechanism of phoneme classification
  - mechanism of lexical recall
  - mechanism of grammar constraints
- ASR is a practical implementation (?)

Outline

Auditory organization

- Detection model is huge simplification
- The real role of hearing is much more general:
  Recover useful information from the outside world

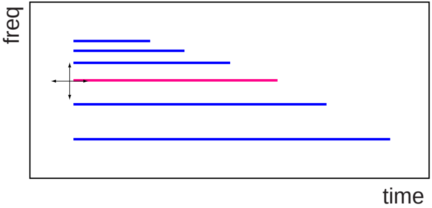
→ Sound organization into events and sources



- Research questions:
  - what determines perception of sources?
  - how do humans separate mixtures?
  - how much can we tell about a source?

Auditory scene analysis: simultaneous fusion

- Harmonics are distinct on AN,
  but perceived as one sound ("fused")

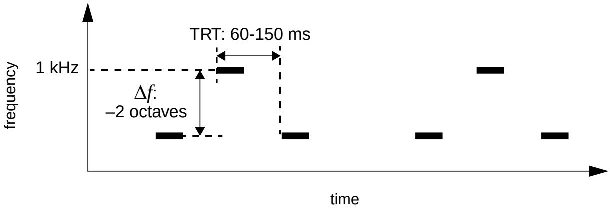


  - depends on common onset
  - depends on harmonicity (common period)

- Methodologies:
  - ask subject how many 'objects'
  - match attributes *e.g.* object pitch
  - manipulate high level *e.g.* vowel identity

Sequential grouping: streaming

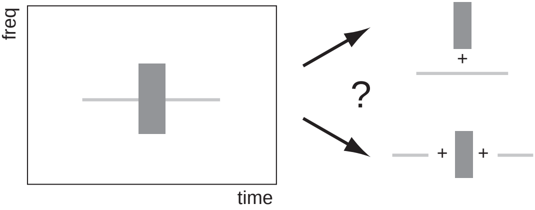
- Pattern / rhythm: property of a set of objects
  - subsequent to fusion ∵ employs fused events?



- Measure by relative timing judgments
  - cannot compare between streams
- Separate 'coherence' and 'fusion' boundaries
- Can interact and compete with fusion

Continuity and restoration

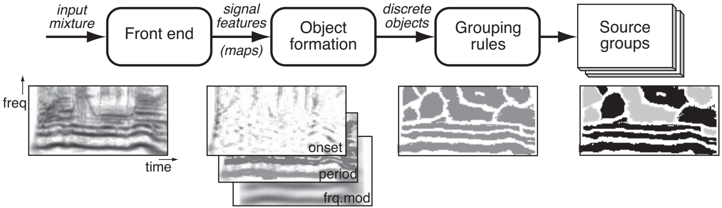
- Tone is interrupted by noise burst: what happened?



  - masking makes tone undetectable during noise
- Need to infer most probable real-world events
  - observation equally likely for either explanation
  - prior on continuous tone much higher $\Rightarrow$ choose
- Top-down influence on perceived events...

Models of auditory organization

Psychological accounts suggest bottom-up



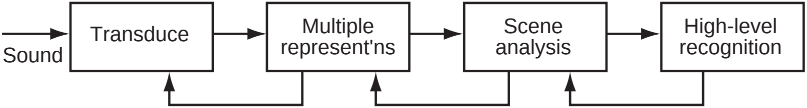
- Brown and Cooke (1994)

Complicated in practice

- formation of separate elements
- contradictory cues
- influence of top-down constraints (context, expectations, . . .)

Summary

- Auditory perception provides the 'ground truth' underlying audio processing
- Physiology specifies information available
- Psychophysics measure basic sensitivities
- Sounds sources require further organization
- Strong contextual effects in speech perception



Parting thought

Is pitch central to communication? Why?

References

Richard M. Warren. Perceptual restoration of missing speech sounds. *Science*, 167 (3917):392–393, January 1970.

R. E. Remez, P. E. Rubin, D. B. Pisoni, and T. D. Carrell. Speech perception without traditional speech cues. *Science*, 212(4497):947–949, May 1981.

G. J. Brown and M. Cooke. Computational auditory scene analysis. *Computer Speech & Language*, 8(4):297–336, 1994.

Brian C. J. Moore. *An Introduction to the Psychology of Hearing*. Academic Press, fifth edition, April 2003. ISBN 0125056281.

James O. Pickles. *An Introduction to the Physiology of Hearing*. Academic Press, second edition, January 1988. ISBN 0125547544.